See Detail Say Clear: Towards Brain CT Report Generation via Pathological Clue-driven Representation Learning

Chengxin Zheng[1], Junzhong Ji[1], Yanzhao Shi[1], Xiaodan Zhang[1,*], Liangqiong Qu[2]

[1]College of Computer Science, Beijing University of Technology, Beijing, China
[2]Department of Statistics and Actuarial Science, School of Computing and Data Science,
The University of Hong Kong, Hong Kong, China
Correspondence: zhangxiaodan@bjut.edu.cn

Abstract

Brain CT report generation is significant to aid physicians in diagnosing cranial diseases. Recent studies concentrate on handling the consistency between visual and textual pathological features to improve the coherence of report. However, there exist some challenges: 1) **Redundant visual representing**: Massive irrelevant areas in 3D scans distract models from representing salient visual contexts. 2) **Shifted semantic representing**: Limited medical corpus causes difficulties for models to transfer the learned textual representations to generative layers. This study introduces a Pathological Clue-driven Representation Learning (PCRL) model to build cross-modal representations based on pathological clues and naturally adapt them for accurate report generation. Specifically, we construct pathological clues from perspectives of segmented regions, pathological entities, and report themes, to fully grasp visual pathological patterns and learn cross-modal feature representations. To adapt the representations for the text generation task, we bridge the gap between representation learning and report generation by using a unified large language model (LLM) with task-tailored instructions. These crafted instructions enable the LLM to be flexibly fine-tuned across tasks and smoothly transfer the semantic representation for report generation. Experiments demonstrate that our method outperforms previous methods and achieves SoTA performance. Our code is available at https://github.com/Chauncey-Jheng/PCRL-MRG.

1 Introduction

Brain computed tomography (CT) imaging is essential for diagnosing various cranial diseases, including cerebral infarction and hemorrhage. However, it is time-consuming and error-prone for radiologists to manually interpret medical findings from these scans and write reports. Automated report generation systems are designed to boost efficiency, reduce the workload for radiologists, and optimize resources in busy clinical scenarios.

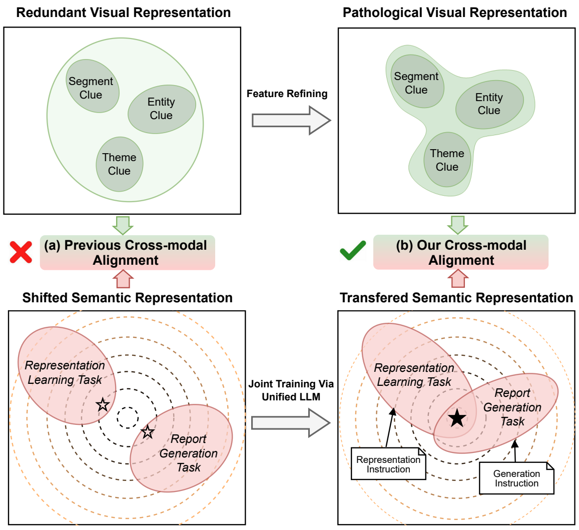

Redundant Visual Representation
Segment Clue
Entity Clue
Theme Clue
Feature Refining
Pathological Visual Representation
Segment Clue
Entity Clue
Theme Clue
(a) Previous Cross-modal Alignment
(b) Our Cross-modal Alignment
Shifted Semantic Representation
Representation Learning Task
Report Generation Task
Joint Training Via Unified LLM
Transfered Semantic Representation
Representation Learning Task
Report Generation Task
Representation Instruction
Generation Instruction

Figure 1: Comparison of different cross-modal alignment paradigms. **(a) Previous way**: The origin CT images contain extraneous information unrelated to diagnosis, and separate training for two semantic tasks makes it challenging to find a shared optimal solution, resulting in inadequate cross-modal alignment. **(b) Our way**: The visual representation is refined to concentrate on pathological clues, and employ joint training with task-tailored instructions via unified LLM to find a transferable representation, leading to better adaption for report generation.

With the advancement of deep neural networks and their successful application in image captioning tasks (Vinyals et al., 2015; Xu et al., 2015), medical report generation (MRG) has gained more attention. Unlike the short sentences of traditional image captioning, MRG aims to generate lengthy and precise reports. To achieve this, various cross-modal alignment methods are required to ensure the consistency between visual and textual information (Shi et al., 2023), including attention mechanisms (Jing et al., 2018; Wang et al., 2018), memory mechanisms (Chen et al., 2020, 2021), and knowledge graphs (Li et al., 2019, 2023b).

---

*Corresponding Author

Recently, learning representations via visual-textual contrastive learning (Shi et al., 2023, 2024) or using pre-trained large language models (LLMs) (Thawakar et al., 2023) to strength representations are also proven to be effective.

However, as shown in Figure 1(a), learning cross-modal correspondences is still challenging in current methods due to the following concerns: 1) **Redundant visual representing.** Different from chest X-ray data, 3D brain CT scans contain extensive redundant information, e.g. background and insignificant areas. With the lack of human-crafted boxes to locate pathology regions, models struggle to capture and interpret the visual pathology patterns for generating reports. Although current advanced models use semantic prior knowledge or medical prompts (Jin et al., 2024; Bu et al., 2024) to automatically learn the salient visual areas, this may introduce noise and unstable representation and cause severe hallucinations in generated texts. 2) **Shifted semantic representing.** Compared to natural corpus, limited brain CT report corpus is insufficient to transfer the pathological semantic representations learned by represent learning layers (e.g., contrastive learning layer) to the language model (Huh et al., 2024), since the direct weight-sharing (Shi et al., 2023, 2024) is prone to degrade the coherence of generated diagnostic sentences. Thus, how to uniformly represent cross-modal pathological features and adapt them to report generation still remains an open question.

In this paper, we propose a Pathological Clue-driven Representation Learning (PCRL) model to seamlessly build cross-modal representations based on diverse pathological clues and transfer them for generating accurate brain CT reports. Specifically, we extract pathological clues from perspectives of segmented regions, pathological entities, and report themes to depict clinical scenarios. Segmented region clues are automatically generated and filtered by given pathology prompts, enabling the visual encoder to grasp visual pathological patterns. Meanwhile, entity and theme clues are respectively extracted by detailed findings and full-text reports, to handle the enriched visual-textual alignment and build cross-modal pathology representations.

Besides, to adapt the learned representations for the text generation task, we bridge the gap between representation learning and report generation by employing a unified large language model (LLM) with task-tailored instructions, which has proven to be more effective than conventional decoders by using appropriate tokens to seamlessly connect different tasks (Yu et al., 2023). As shown in Figure 1(b), We craft a representation instruction to prompt the LLM to produce high-level pathological semantic features for cross-modal alignment, and a generation instruction to prompt LLM to generate accurate brain CT reports based on the learned representations.

Our main contributions can be summarized as:

1. We propose a novel framework to seamlessly learn visual-textual representations from perspectives of diverse pathological clues and leverage them for enhancing the quality of generated brain CT reports.
2. We, for the first time, design a new paradigm to effectively transfer the learned pathological representations for report generation using a unified LLM prompted by task-tailored instructions.
3. We validate the model capabilities on the open-source CTRG-Brain dataset. Experimental results show that our model achieves remarkable performance in generating brain CT reports.

2 Related Work

2.1 Medical Report Generation

The advancements in image captioning techniques have spurred the development of a range of radiology report generation methods (Jing et al., 2018; Chen et al., 2020, 2021; Li et al., 2019; Shi et al., 2023; Zhang et al., 2023; Li et al., 2023b; Shi et al., 2024; Shen et al., 2024). To exploit the more effective parts of medical images, Jing et al. (2018) first propose a co-attention mechanism to associate images with disease tags and improve the accuracy of generated reports; Chen et al. (2021) design a cross-modal memory matrix to learn high-level vision-language correspondence for enhancing report generation; Liu et al. (2021c) used contrastive attention to captured the difference of abnormal and normal samples. To address the semantic bias of limited medical report corpus, Jing et al. (2020) exploited the textual structure information of reports, while Liu et al. (2021a); Shen et al. (2024) used curriculum learning to alleviate the data bias of limited medical data corpus. The retrieval-based

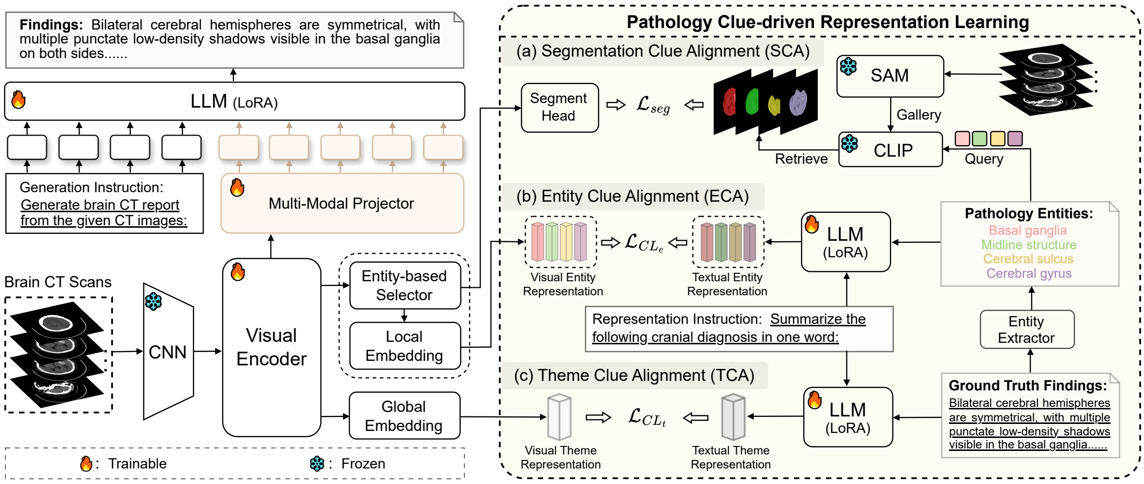

Findings: Bilateral cerebral hemispheres are symmetrical, with multiple punctate low-density shadows visible in the basal ganglia on both sides......
LLM (LoRA)
Generation Instruction: Generate brain CT report from the given CT images:
Multi-Modal Projector
Brain CT Scans
CNN
Visual Encoder
Entity-based Selector
Local Embedding
Global Embedding
: Trainable
: Frozen
Pathology Clue-driven Representation Learning
(a) Segmentation Clue Alignment (SCA)
Segment Head
$\mathcal{L}_{seg}$
SAM
Gallery
CLIP
Retrieve
Query
(b) Entity Clue Alignment (ECA)
$\mathcal{L}_{CL_e}$
Visual Entity Representation
Textual Entity Representation
LLM (LoRA)
Pathology Entities:
Basal ganglia
Midline structure
Cerebral sulcus
Cerebral gyrus
Representation Instruction: Summarize the following cranial diagnosis in one word:
Entity Extractor
(c) Theme Clue Alignment (TCA)
$\mathcal{L}_{CL_t}$
Visual Theme Representation
Textual Theme Representation
LLM (LoRA)
Ground Truth Findings: Bilateral cerebral hemispheres are symmetrical, with multiple punctate low-density shadows visible in the basal ganglia......

Figure 2: The overall framework of our method, which mainly consists of an image encoder and a text decoder for brain CT report generation (left). The pathological clue-driven representation learning (right) is proposed to guide the encoder and decoder for more fine-grained representation by three alignment modules: (a) segmentation clue alignment (SCA), (b) entity clue alignment (ECA), and (c) theme clue alignment (TCA).

(Li et al., 2018; Liu et al., 2021b) and prior expert knowledge based methods (Zhang et al., 2020; Yang et al., 2022; You et al., 2021) also shows effectiveness to alleviate the inaccuracy of generated report caused by limited training data. To build more stable relations between multi-modal features, contrastive learning (Radford et al., 2021) acts as an unsupervised method, and is proven to be effective for learning cross-modal representations for MRG tasks (Li et al., 2023b; Shi et al., 2024).

Recently, LLM-based MRG methods (Jin et al., 2024; Chen et al., 2024) harness the semantic processing capabilities of LLMs to enhance the coherence of generated medical reports. However, the current bottleneck of LLM is to interpret sparse visual patterns of pathologies, which causes severe hallucinations and degrades the trustworthiness of clinical diagnosis. To solve this issue, we propose to augment LLM with representation learning guided by pathology clues. By aligning cross-modal pathological features and transferring the learned representations to the LLM-based report decoder, our model can generate accurate and coherent brain CT reports.

2.2 Pre-trained Large Model

In recent years, pre-trained large models have achieved significant breakthroughs in both natural language processing (NLP) and computer vision (CV). These models leverage extensive datasets for pre-training, allowing them to perform exceptionally well in downstream tasks. The Segment Anything Model (SAM) Kirillov et al. (2023) is an innovative large vision model in the field of image segmentation. SAM's strengths lie in its extensive training on a large-scale dataset, enhancing its versatility and enabling real-time generation of high-quality segmentation masks across diverse tasks and image distributions. MedCLIP Wang et al. (2022b) is a notable visual-language model in the medical domain. Pre-trained on decoupling medical images and texts, MedCLIP has shown SoTA performance on zero-shot image-text retrieval. LLaMA3 is the next generation of state-of-the-art open-source LLM from Meta (2024). Trained on over 15 trillion tokens, seven times the data of its predecessor, LLaMA3 incorporates extensive multilingual and high-quality datasets. To improve the use of large models for assisting in medical report generation, we adopt SAM for region segmentation, MedCLIP for feature representation of pathological entities, and LLaMA for joint training to achieve fine-grained cross-modal alignment and improve medical report generation.

3 Methodology

As illustrated in Figure 2, our framework consists of two branches: the brain CT Report Generation (RG) on the left and the Pathological Clue-driven Representation Learning (PCRL) on the right. The interaction between these branches is facilitated through the shared use of a visual encoder and a language model.

3.1 Brain CT Report Generation

In this branch, the input comprises a set of CT scan images $I = \{i_1, \ldots, i_N\}$, where N represents the number of CT images in each sample. The output is the corresponding brain CT findings report $Y = \{y_1, \ldots, y_M\}$, with M representing the number of tokens. We adopt the encoder-decoder architecture for report generation. First, we use ResNet101 (He et al., 2016) to extract grid features $G = \{g_1, \ldots, g_N\} \in \mathbf{R}^{N \times H \times d} (N = 24, H = 196, d = 2048)$ from I. Then, we embed these features using a visual encoder to obtain visual features V. These visual features V are passed through a multi-modal projector to obtain the visual embedding tokens $H_v \in \mathbf{R}^{N \times d_w} (d_w = 4096)$, which are aligned with the word embedding space of a large language model. Finally, these tokens are concatenated with the instruction token H_q and fed into the large language model. We train the parameters θ by minimizing the cross-entropy loss, which can be represented by the following formula:

$$\mathcal{L}_g = -\sum_{t=1}^{M} \log P(y_t \mid y_{1:t-1}, H_v, H_q; \theta) \quad (1)$$

where, $P(y_t|*)$ denotes the probability conditioned on H_v, H_q, and the embeddings of previous words $y_1, y_2, \ldots, y_{t-1}$.

3.2 Pathological Clue-driven Representation Learning

Learning fine-grained visual-text representations is critical for generating accurate reports (Shi et al., 2023). This branch mainly builds enriched representations by conducting multi-modal feature alignment based on pathological clues, including segmentation clue alignment (SCA), entity clues alignment (ECA), and theme clues alignment (TCA).

3.2.1 Preparation of Pathological Clues

Pathological clues are collected in three perspectives for feature alignment.

Pathology Theme Clues: Learning the structure of professional brain CT reports is essential for MRG models to satisfy human standards and produce reliable reports. To this end, we propose to build theme clues by simply using the full-text report and whole images as global signals, which can be useful to enhance the overall quality of medical reports by TCA (illustrated in 3.2.4).

Pathology Entity Clues: To learn detailed information about pathology entities, we regard each single finding sentence in the report as an entity clue and extract the related multi-modal features in ECA 3.2.3. The construction of entities $E = \{e_1, ..., e_{N_e}\}$ (with $N_e = 24$) is based on expert knowledge and the frequency of words in the training corpus.

Segment Clues: To enhance visual representations via detailed contour of pathologies, we propose to utilize the pre-trained segmentation model SAM (Kirillov et al., 2023) to generate mask candidates and filter useful masks related to pathology entities. First, we prompt SAM by covering each brain CT image with a grid of points and integrating it into image embeddings through average sampling. In this way, the mask decoder in SAM is prompted to generate a gallery of candidate masks $M_I = \{m_1, ..., m_{N_m}\}$. To ensure the quality of masks, we also apply rule-based methods to filter out low-quality and duplicate masks with area size, stability scores, and IoU scores. For each sample consisting of 24 images, we generate a corresponding series of masks and combine them to form the candidate $Gallery = \{M_{I_1}, ..., M_{I_N}\}$.

Then, to retrieve valuable masks related to the sample's pathological entities, we propose to use the MedCLIP (Wang et al., 2022b) for text-prompted retrieval. Based on expert knowledge, we divide the 24 images into eight-layer categories, each mapping a specific set of entities (Shi et al., 2024). Conversely, each entity also has its corresponding layers. We extract the existing entities from the sample's report and their corresponding pathological descriptions $D = \{d_1, ..., d_{N_d}\}$ ($N_d \leq N_e$), which serve as the $Query$. Next, we use the d-th description $Query_d$ to search for the most similar matching mask in $Gallery_d$, which is a subset of $Gallery$, consisting of all visual masks of the layers related to the existing entities. The procedure can be represented as:

$$Retrieval_d = CLIP(Gallery_d, Query_d) \quad (2)$$

where $Retrieval_d$ denotes the retrieved masks for SCA 3.2.2.

3.2.2 Segmentation Clue Alignment

This module aims to learn the fine-grained visual representation based on the extracted segment masks. First, we obtain the visual features $V_D = \{v_1, ...v_{N_d}\}$ corresponding to the images containing the entities using a selector. Then, we input these features into a lightweight segmentation head to generate the corresponding foreground entity masks $S_D = \{s_1, ..., s_{N_d}\}$, which serves as discriminated pathological information. We then align these generated masks with the retrieved segment masks $M_D = \{Retrieval_1, ..., Retrival_{N_d}\}$ to learn detailed patterns. It is important to note that the size of the generated masks S_D differs from the size of the retrieved SAM masks M_D. To address this, we first resize the SAM masks to match the size of the masks generated by our segmentation head. We finally calculate the following loss for aligning the two types of masks:

$$\mathcal{L}_{seg} = 1 - \frac{2\sum_{i=1}^{n} p_i y_i}{\sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} y_i^2} \tag{3}$$

where y_i is the pixel value of retrieved SAM masks and p_i is the predicted pixel value.

In this way, the visual encoder can be effectively learned to focus on the areas of pathologies while reducing the influence of irrelevant visual features.

3.2.3 Entity Clue Alignment

To learn the cross-modal patterns of pathology entities, we extract visual and textual pathological features based on the entity clues. First, we build a selector to obtain the visual features $V_D = \{v_1, ...v_{N_d}\}(N_d < N)$, which corresponds to significant CT images that exist pathology entities. We then apply global average pooling (GAP) to generalize V_D and obtain representations of each significant CT image, denoted as $R_v^e = \{r_1, ...r_{N_d}\}$.

Different from previous work (Li et al., 2023b) use an external language model (e.g. SciBert (Beltagy et al., 2019)) to build textual representation for feature alignment, we propose to leverage a unified LLM to generate textual representation via tailored prompts. Inspired by Wang et al. (2023), we crafted a representation instruction "*Summarize the following cranial diagnosis in one word:*", which is denoted as H_{q_r}. This prompt can effectively activate the summarization ability of LLM to generate corresponding text representations for pathological entity description in D. We use the output generated by the final hidden layer of LLM as textual features for cross-modal alignment.

With the carefully extracted visual and textual representations of entities, we map them into the same dimension through an embedding layer. The process can be represented by the following formula:

$$\begin{aligned} R_v^e &= Linear_v^e(GAP(V_D)) & (4) \\ R_w^e &= Linear_t^e(LLM(H_{q_r}, D)) & (5) \end{aligned}$$

where GAP denotes the global average pooling, $Linear_v$ represents the visual mapper, and $Linear_t$ represents the textual mapper. We employ the symmetric InfoNCE (van den Oord et al., 2018) loss for visual-textual alignment:

$$\begin{aligned} \mathcal{L}_{CL_e} = -\frac{1}{2}\Bigg(&\alpha_v \sum_{i=1}^{N_d} \log \frac{\exp(\mathrm{s}_v^e(i,i))}{\sum_{j=1,j\neq i}^{N_d} \exp(\mathrm{s}_v^e(i,j))} \\ &+ \alpha_w \sum_{i=1}^{N_d} \log \frac{\exp(\mathrm{s}_w^e(i,i))}{\sum_{j=1,j\neq i}^{N_d} \exp(\mathrm{s}_w^e(i,j))}\Bigg) \end{aligned} \tag{6}$$

where $s_v^e(i,j) = sim(R_{v_i}^e, R_{w_j}^e)/\tau$ and $s_w^e(i,j) = sim(R_{w_i}^e, R_{v_j}^e)/\tau$ denote the similarity between the visual representation R_v^e and the textual representation R_w^e, τ is a temperature parameter. α_v and α_w are hyperparameters to balance the contrastive learning.

By aligning multi-modal entity clues, the model can grasp fine-grained visual-text representations to generate accurate diagnostic words.

3.2.4 Theme Clue Alignment

Theme clue alignment aims to equip the model with comprehensive skills to generate accurate style and structure of reports, thereby enhancing clinical reliability. For global visual features, we use one-dimensional global max pooling (GMP) to represent the entire sample visually, denoted as R_v^t. For the overall report of the sample, we generate textual representation R_w^t by prompting LLM with the same representation instruction (see Section 3.2.3). This process can be represented as:

$$\begin{aligned} R_v^t &= Linear_v^t(GMP(V)) & (7) \\ R_w^t &= Linear_t^t(LLM(H_{q_s}, Y)) & (8) \end{aligned}$$

Similar to ECA, the loss of TCA can be formulated as:

$$\begin{aligned} \mathcal{L}_{CL_t} = -\frac{1}{2}\Bigg(&\alpha_v \sum_{i=1}^{N_b} \log \frac{\exp(\mathrm{s}_v^t(i,i))}{\sum_{j=1,j\neq i}^{N_b} \exp(\mathrm{s}_v^t(i,j))} \\ &+ \alpha_w \sum_{i=1}^{N_b} \log \frac{\exp(\mathrm{s}_w^t(i,i))}{\sum_{j=1,j\neq i}^{N_b} \exp(\mathrm{s}_w^t(i,j))}\Bigg) \end{aligned} \tag{9}$$

where $s_v^t(i,j) = sim(R_{v_i}^t, R_{w_j}^t)/\tau$ and $s_w^t(i,j) = sim(R_{w_i}^t, R_{v_j}^t)/\tau$ denote the similarity between the visual representation R_v^e and the textual representation R_w^e, τ is a temperature parameter, N_b is the batch size. Shared with ECA, α_v and α_w are set to balance the feature alignment.

Methods	Decoder	B1	B2	B3	B4	M	RG	C	F1
HRNN(Krause et al., 2017)†	LSTM	42.3	28.3	21.2	17.1	27.2	39.3	20.9	70.2
Up-Down(Anderson et al., 2018)†	LSTM	45.8	34.8	28.5	24.4	31.6	42.5	27.3	70.2
WCL(Yan et al., 2021)†	LSTM	49.5	36.5	29.4	25.1	31.3	42.8	33.3	64.5
R2Gen-CMN(Chen et al., 2021)†	Transformer	49.1	40.0	34.4	30.1	29.9	48.6	84.2	69.8
XProNet(Wang et al., 2022a)†	Transformer	50.6	41.3	34.4	29.1	31.5	51.7	83.3	70.1
WGAM(Yang et al., 2021)†	LSTM	49.4	36.7	29.6	25.4	32.0	42.4	31.9	68.9
PGCA(Shi et al., 2023)†	LSTM	50.2	37.8	30.7	26.5	32.5	43.0	34.0	69.2
WGAM-HI(Zhang et al., 2023)†	LSTM	50.4	37.6	30.5	26.1	31.4	43.8	33.2	67.4
LLaVA-med(Li et al., 2023a)†	LLaMA3-8B	50.0	39.3	32.0	26.3	31.4	46.7	38.6	64.3
HILT(Liu et al., 2024)†	LLaMA3-8B	50.7	39.9	33.1	27.9	30.8	46.1	43.7	68.4
PromptMRG(Jin et al., 2024)†	LLaMA3-8B	48.1	38.3	31.6	26.5	31.0	47.4	50.3	69.2
Ours (Jieba tokenizer)	LLaMA3-8B	51.5	42.0	35.7	30.9	31.4	49.0	80.0	**70.6**
Ours (LLaMA tokenizer)	LLaMA3-8B	**62.0**	**54.7**	**49.4**	**45.3**	**33.1**	**57.7**	**96.4**	**70.6**

Table 1: The performance of our PCRL compared with previous state-of-the-art models on the Brain CT report generation dataset *CTRG-Brain*. The best results are highlighted in bold. † denotes the re-implementation results.

3.3 Joint Training

In the training stage, we jointly train the RG branch and the PCRL branch to maximize the utilization of multi-granularity visual-text representations. Instead of using separate modules to learn representation and generate medical reports (Li et al., 2023b), we use a unified LLM to bridge the gap between representation learning and report generation via two task-tailored instructions, i.e., representation instruction and generation instruction. This can transfer the representations learned by the PCRL branch to optimize the RG branch effectively, thereby generating accurate reports.

Our final loss contains the above-mentioned losses, which can be formulated as:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_{seg} + \mathcal{L}_{CL_e} + \mathcal{L}_{CL_t} \tag{10}$$

4 Experiments

4.1 Dataset

We validate the performance of our model using the *CTRG-Brain* (Tang et al., 2024) dataset. This dataset comprises 6,000 samples, containing a total of 160,336 CT images and 6,000 Chinese medical reports. Following the mainstream division method (Shi et al., 2023; Zhang et al., 2023), we split the dataset into a training set, a validation set, and a test set in a 7:1:2 ratio. For consistent processing, we divide the brain CT image samples into 8 layers based on expert knowledge, with each layer containing 3 continuous CT images, assigning each sample with 24 CT images.

4.2 Evaluation Metrics

We chose Natural Language Generation (NLG) metrics and Clinical Evaluation (CE) metrics to evaluate our model's performance. NLG metrics include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), denoted as B1, B2, B3, B4, M, RG, and C respectively. To measure the pathological accuracy, we use 24 keywords summarized by experienced radiologists to calculate the Clinical Evaluation (CE) metric(Shi et al., 2023; Zhang et al., 2023), i. e., F1 score, which is the harmonic mean of the precision and recall.

4.3 Implementation Details

We reshape the size of image to 512x512 pixels and used a ResNet101 to extract image features, which is pre-trained on the ImageNet dataset and fine-tuned on the CQ500 dataset (Chilamkurthy et al., 2018). For our large language model, we utilize LLaMA3-8B (Meta, 2024), which is quantized to 4-bit, and use the LoRa (Hu et al., 2021) for parameter-efficient fine-tuning. The overall trainable parameter quantity of our model is 229.9M, with 3.4M parameters in LLM (only 0.04%). During training, we use the AdamW optimizer with a learning rate of 1e-4. The batch size is 4, with 1050 training steps per epoch. For testing, we set the temperature coefficient of the large model to 0.6 and the top-p value to 0.9. The model is implemented using PyTorch 2.3.0, and the entire training process is conducted on a single RTX 4090 GPU.

Methods	Representation		Module Loss			B1	B2	B3	B4	M	RG	C
	Visual	*Textual*	$\mathcal{L}_{SCA}$	$\mathcal{L}_{ECA}$	$\mathcal{L}_{TCA}$							
Baseline	✗	✗	✗	✗	✗	49.6	39.9	33.4	28.6	30.7	48.4	50.8
(a)	✗	✗	✓	✗	✗	51.3	41.0	34.5	29.7	31.2	47.7	70.8
(b)	✓	✓	✗	✓	✗	51.3	41.4	35.0	30.1	31.2	48.7	71.3
(c)	✓	✓	✓	✓	✗	**51.7**	41.4	34.6	29.5	31.2	48.0	72.8
(d)	✓	✓	✗	✓	✓	48.9	38.8	32.2	27.4	30.0	46.3	58.3
(e)	✓	✓ (Bert)	✓	✓	✓	49.5	39.0	31.8	26.3	30.4	45.4	48.5
Ours	✓	✓	✓	✓	✓	51.5	**42.0**	**35.7**	**30.9**	**31.4**	**49.0**	**80.0**

Table 2: Ablation studies of our proposed method. The **Baseline** model is an encoder-decoder framework without an alignment mechanism. **(a)**, **(b)**, **(c)**, **(d)** and **(e)** respectively denote the use of different representations (i.e., visual and textual in ECA and TCA) and module losses.

4.4 Quantitative Analysis

We compare the proposed PCRL with some competitive brain CT report generation methods (**WGAM** (Yang et al., 2021), **PGCA** (Shi et al., 2023), **WGAM-HI**(Zhang et al., 2023)). Besides, we also reproduced some SOTA models in image captioning (**HRNN** (Krause et al., 2017), **Up-Down** (Anderson et al., 2018)) and chest X-ray report generation (**WCL** (Yan et al., 2021), **R2Gen-CMN** (Chen et al., 2021), **XProNet** (Wang et al., 2022a)) for comprehensive comparisons on *CTRG-Brain* dataset. What's more, for fair comparisons, We also reproduce several related LLM-based methods (**LLaVA-Med**(Li et al., 2023a), **HILT**(Liu et al., 2024), **PromptMRG**(Jin et al., 2024)with the same LLM decoder.

As shown in Table 1, our method outperforms others across most evaluation metrics. WGAM and WGAM-HI employ weakly-supervised visual attention to extract key visual features, resulting in higher BLEU scores. With contrastive learning for cross-modal alignment, WCL and PGCA can effectively learn relations between CT images and reports, achieving better results. However, due to the lack of training data, the above methods still produce reports that fall short in fluency and readability. Here's a refined version of your sentence: The LLM-based methods (LLaVA-med, HILT, PromptMRG) demonstrate poorer performance compared to traditional transformer-based methods (R2Gen-CMN, XProNet). This suggests that prior knowledge from pretrained large models, without further cross-modal alignment, may lead to hallucinations in report generation.

Our PCRL achieves fine-grained cross-modal alignment by leveraging a series of pre-trained large models, generating more fluent reports and achieves the best performance compared to other methods. It is noteworthy that we tested our method using the Jieba and LLaMA tokenizers respectively to compute NLG scores, with the latter achieving the best performance across all NLG metrics.

This discrepancy may be due to differences in tokenization methods used during training. While other models employ the traditional Jieba tool for Chinese word tokenization, our PCRL follows more advanced BPE-based subword tokenization. Nevertheless, our model also achieves competitive results in overall metrics.

4.5 Ablation Study

To evaluate the effect of each component in PCRL, we have done plenty of ablation studies, as shown in Table 2. Baseline is the standard encoder-decoder (ResNet101-LLaMA3) architecture without alignment. By progressively adding visual or textual representation (*visual* and *textual*) and the module losses ($\mathcal{L}_{SCA}$, $\mathcal{L}_{ECA}$ and $\mathcal{L}_{TCA}$), respectively denoting the incorporation of two modalities of representations and the utilization of three pathological clue-driven alignments (i.e., SCA, ECA, and TCA).

By comparing (a) with the baseline, we observe that incorporating the segmentation alignment significantly enhances the model's performance. This demonstrates that our SCA method effectively filters out irrelevant information from CT images and extracts crucial pathology-related visual regions. In contrast, with the implementation of ECA, (b) achieves comparable or better performance across all metrics compared to (a). This indicates that fine-grained cross-modal pathological entity alignment can effectively learn strongly correlated visual-text

---

https://github.com/fxsjy/jieba

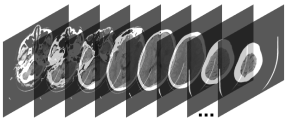
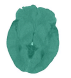
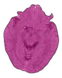
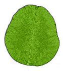
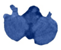
Ground Truth Sample	Ground Truth Entities	Retrieved Masks	Generated Reports	
 双侧大脑半球对称，左侧**基底节区**见小片状低密度影，边缘欠清晰，无占位效应，各**脑室**、脑池大小形态正常，**中线**结构居中，幕下小脑，**脑干**无异常。 (The bilateral cerebral hemispheres are symmetrical. A small patch of low-density shadow is seen in the left basal ganglia region, with unclear edges and no mass effect. The size and shape of all ventricles and cisterns are normal. The midline structure is centered. There are no abnormalities in the cerebellum and brainstem below the tentorium.)	**基底节区(basal ganglia) :** 左侧基底节区见小片状低密度影 (Small patchy low-density shadows visible in the left basal ganglia area) **脑室(ventricle):** 各脑室、脑池大小形态正常 (The size and shape of all ventricles and cisterns are normal.) **中线(midline structure):** 中线结构居中 (The midline structure centered) **脑干(brainstem):** 幕下小脑，脑干无异常 (There are no abnormalities in the cerebellum and brainstem below the tentorium)	   	双侧大脑半球对称，两侧**基底节区**见斑点状低密度影，**中线**结构居中，脑沟、脑池、蛛网膜下腔稍增宽，脑回变小。(Bilateral cerebral hemispheres are symmetrical, with punctate low-density shadows in the basal ganglia on both sides. The midline structure is centered, and the sulcus, cistern, and subarachnoid space are slightly widened, while the gyrus is smaller.)	**Baseline**
			双侧大脑半球对称，右侧丘脑见小片状低密度影，边缘欠清晰，无占位效应，各**脑室**、脑池大小形态正常，**中线**结构居中，幕下小脑、**脑干**无异常。(The bilateral cerebral hemispheres are symmetrical. A small patch of low-density shadow is seen in the right thalamus, with unclear edges and no mass effect. The size and shape of all ventricles and cisterns are normal. The midline structure is centered. There are no abnormalities in the cerebellum and brainstem below the tentorium.)	**WGAM-HI**
			双侧大脑半球对称，左侧**基底节区**见斑片状低密度影，边缘欠清晰，无占位效应，各**脑室**、脑池大小形态正常，**中线**结构居中，幕下小脑，**脑干**无异常。 (The bilateral cerebral hemispheres are symmetrical. A patch of low-density shadow is seen in the left basal ganglia region, with unclear edges and no mass effect. The size and shape of all ventricles and cisterns are normal. The midline structure is centered. There are no abnormalities in the cerebellum and brainstem below the tentorium.)	**Ours**

Figure 3: Visualization of report generation and mask segmentation. Given the ground truth sample and corresponding entities, the retrieved entity masks are listed in the third column. Reports generated by Baseline, WGAM-HI, and our model are listed in the fourth column. Different colors denote the specific entity words and entity masks, respectively. The English translation is given for a better understanding of the original Chinese reports in CTRG-Brain.

representations, thus better supporting medical report generation. Besides, (c) combines both SCA and ECA and shows improved performance in the B1 and C metrics, which are highly correlated with keyword frequency. However, it slightly underperforms in the B2, B4 and RG metrics compared to (b), which is more related to overall text style.(d) showed SCA is essential for focusing on entity details and maintaining the thematic style.

Notably, to validate the contribution of joint representation learning, we replace the language model with Bert model for textual representation during alignment in (e). The result indicated that shared representation through joint learning with unified language model can effectively improve the overall quality of the task.

4.6 Qualitative Analysis

We visualize the brain CT reports generated by baseline, WGAM-HI (Zhang et al., 2023) and our model in Figure 3. Given the ground truth brain CT sample and entity data, our model generates better brain CT report with the most accurate entity words (e.g. basal ganglia, ventricle, and brainstem) among the competitors, which demonstrates the effectiveness of using diverse medical clues to learn enriched multi-modal representations. It can be also seen that compared with the baseline, the report generated by our model has a better semantic structure, indicating the contribution of employing a unified LLM to transfer useful representations. Besides, we also find that the retrieved entity masks from segmentation masks gallery (see Figure 4) can generally match related entity words at both levels of visual and semantic. For example, the mask of "midline structure" and "brainstem" matches the empirical scan slice and fine-grained region chosen by experienced radiologists. This guarantees the model to mine accurate visual cranial patterns, therefore generating high-quality reports.

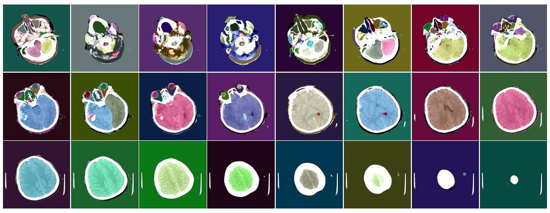


Figure 4: Visualization of the segmentation masks gallery generated by SAM for a single sample.

5 Conclusion

We propose a novel model to mine pathological clues for enhancing multi-modal representations and seamlessly transfer them into report generation. First, through carefully designed segmentation clue alignment, entity clue alignment, and theme clue alignment, the diverse and precise feature representation can be well-constructed. Second, we transfer the learned representation to boost the brain CT report generation via a unified LLM prompted by task-tailored instructions. Experiments demonstrate the effectiveness of our model in generating pathologically accurate reports.

Limitations

Although the segmentation clues retrieved by MedClip (Wang et al., 2022b) can generally match corresponding pathological entities and help the model neglect redundant visual information, it should be noted that a part of retrieved entity masks may not be precise. This is because MedClip is mainly pretrained by chest X-ray data with limited brain CT samples. Thus, addressing this challenge is imperative for the research community. In the future, we will work on exploring useful approaches. One potential approach is to train a unified text-prompted medical segmentation model towards 3D brain CT scans, which can not only be employed to offer fine-grained visual information for medical report generation but also for other related tasks, e.g. medical VQA.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086.

S Banerjee and A Lavie. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*, pages 65–72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618.

Shenshen Bu, Yujie Song, Taiji Li, and Zhiming Dai. 2024. Dynamic knowledge prompt for chest x-ray report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING*, pages 5425–5436. ELRA and ICCL.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1439–1449.

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. 2024. Dia-llama: Towards large language model-driven CT report generation. *CoRR*, abs/2403.16386.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. Development and validation of deep learning algorithms for detection of critical findings in head ct scans. *arXiv preprint arXiv:1803.05854*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*, pages 2607–2615.

Baoyu Jing, Zeya Wang, and Eric Xing. 2020. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2577–2586.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3337–3345.

C Li, C Wong, S Zhang, et al. 2023a. Llava-med: training a large language-and-vision assistant for biomedicine in one day. arxiv. *arXiv preprint arXiv:2306.00890*.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6666–6673.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. 2024. Benchmarking and boosting radiology report generation for 3d high-resolution medical images. *arXiv preprint arXiv:2406.07146*.

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3001–3012.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13753–13762.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. 2021c. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763.

Qingya Shen, Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Huimin Xu. 2024. Ghcl: Gaussian heuristic curriculum learning for brain ct report generation. *Multimedia Systems*, 30(2):69.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Ying Liu, Zheng Wang, and Huimin Xu. 2024. Prior tissue knowledge-driven contrastive learning for brain CT report generation. *Multim. Syst.*, 30(2):98.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu, and Ying Liu. 2023. Granularity matters: Pathological graph-driven cross-modal alignment for brain CT report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 6617–6630.

Yuhao Tang, Haichen Yang, Liyan Zhang, and Ye Yuan. 2024. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237:121442.

Omkar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *CoRR*, abs/2306.07971.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, pages 563–579. Springer.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *ArXiv*, abs/2401.00368.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9049–9058.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, Julian J. McAuley, and Chun-Nan Hsu. 2021. Weakly supervised contrastive learning for chest x-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4009–4015.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

Sisi Yang, Junzhong Ji, Xiaodan Zhang, Ying Liu, and Zheng Wang. 2021. Weakly guided hierarchical encoder-decoder network for brain ct report generation. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, December 9-12, 2021*, pages 568–573. IEEE.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III*, pages 72–82.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. 2023. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*.

Xiaodan Zhang, Sisi Yang, Yanzhao Shi, Junzhong Ji, Ying Liu, Zheng Wang, and Huimin Xu. 2023. Weakly guided attention model with hierarchical interaction for brain ct report generation. *Computers in Biology and Medicine*, page 107650.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12910–12917.